The LINNEAN SOCIETY of London BIOLOGICAL Journal of the Linnean Society

Biological Journal of the Linnean Society, 2015, ••, ••–••. With 4 figures

First steps in studying the origins of secondary woodiness in *Begonia* (Begoniaceae): combining anatomy, phylogenetics, and stem transcriptomics

CATHERINE KIDNER[1,2], ANDREW GROOVER[3,4], DANIEL C. THOMAS[5,6,7], KATIE EMELIANOVA[1,2], CLAUDIA SOLIZ-GAMBOA[5] and FREDERIC LENS[5]*

[1]*Royal Botanic Gardens, Edinburgh, UK*
[2]*Institute of Molecular Plant Sciences, University of Edinburgh, Edinburgh, UK*
[3]*US Forest Service, Pacific Southwest Research Station, Davis, CA, USA*
[4]*Department of Plant Biology, UC Davis, Davis, CA, USA*
[5]*Naturalis Biodiversity Center, Leiden University, P.O. Box 9517, 2300RA Leiden, The Netherlands*
[6]*School of Biological Sciences, University of Hong Kong, Hong Kong*
[7]*Research and Conservation Branch, Singapore Botanic Gardens, Singapore*

Received 14 November 2014; revised 19 December 2014; accepted for publication 19 December 2014

Since Darwin's observation that secondary woodiness is common on islands, the evolution of woody plants from herbaceous ancestors has been documented in numerous angiosperm groups. However, the evolutionary processes that give rise to this phenomenon are poorly understood. To begin addressing this we have used a range of approaches to study the anatomical and genetic changes associated with the evolution and development of secondary woodiness in a tractable group. *Begonia* is a large, mainly herbaceous, pantropical genus that shows multiple shifts towards secondarily woody species inhabiting mainly tropical montane areas throughout the world. Molecular phylogenies, including only a sample of the woody species in *Begonia*, indicated at least eight instances of a herbaceous–woody transition within the genus. Wood anatomical observations of the five woody species studied revealed protracted juvenilism that further support the secondary derived origin of wood within *Begonia*. To identify potential genes involved in shifts towards secondary woodiness, stem transcriptomes of wood development in *B. burbidgei* were analysed and compared with available transcriptome datasets for the non-woody *B. venustra*, *B. conchifolia*, and *Arabidopsis*, and with transcriptome datasets for wood development in *Populus*. Results identified a number of potential regulatory genes as well as variation in expression of key biosynthetic enzymes.
© 2015 The Linnean Society of London, *Biological Journal of the Linnean Society*, 2015, ••, ••–••.

ADDITIONAL KEYWORDS: molecular wood pathway – wood anatomy.

INTRODUCTION

Charles Darwin was one of the first scientists to mention the occurrence of woodiness in some island species belonging to otherwise herbaceous plant groups (Darwin, 1859). He identified 'insular woodiness' as a derived or *secondary* state, which evolved after ancestral herbaceous progenitor populations reached the islands. For many of these woody island species this shift in habit was later confirmed by molecular phylogenetics, and anatomical studies revealed that the wood in such species was characterized by protracted juvenilism (Carlquist, 1974, 2012; Givnish, 1998; Whittaker & Fernández-Palacios, 2007; Lens *et al*., 2009, 2013a). There is growing evidence that evolutionary shifts towards the woody habit occur convergently within families, on single islands, but also in continental areas with at least some consecutive dry months per year (Carlquist, 1974; Lens *et al*., 2013a; F. Lens, in prep).

*Corresponding author. E-mail: frederic.lens@naturalis.nl

Distinguishing between herbaceousness and woodiness is not always easy due to the continuous variation between both growth forms. Plants from most dicotyledonous angiosperm lineages – including herbaceous groups – retain the ability to produce a limited amount of wood in the basal parts of their stems (Dulin & Kirchoff, 2010; Lens *et al*., 2012a; Lens, Smets & Melzer, 2012b). The limited wood formation in these herbaceous lineages may contribute biomechanical strength to the stem, and could explain the retention of wood forming genes in herbaceous groups. In addition, pleiotropic activity of cambium genes in shoot apical meristem function may also contribute to their retention (Robischon *et al*., 2012; Zhang *et al*., 2014). The potential genetic simplicity of a habit switch from herbaceousness back to the ancestral woody habit in angiosperms is supported in *Arabidopsis* by the 2-gene model of Melzer *et al*. (2008). In this case, loss of function in two MADS box transcription factors enable the herbaceous wild-type to develop into a woody shrub.

Secondary woodiness appears to be correlated with extreme conditions in at least some groups. Drier habitats and competition have been suggested as drivers (Lens *et al*., 2013a, b) and different drivers appear to operate even within the same clade (Lens *et al*., 2009). It has also been suggested that woody growth is a response to more favourable climatic environments (especially lack of frost; Carlquist, 1974) or to promote outcrossing (Böhle, Hilger & Martin, 1996), and that flexible developmental genetics allowing lineages to switch between herbaceous and woody forms may have contributed to the evolutionary success of angiosperms (Bond, 1989). A better understanding of the evolution of secondary woodiness will help us assess the advantages that the woody habit confers.

Genomic and molecular tools are becoming available to study the environmental drivers and proximal genetic causes of this derived wood formation in otherwise herbaceous lineages in a wide array of angiosperms. In particular, transcriptome sequencing can reveal useful information about biosynthetic pathways, regulatory pathways and targets of selection even in non-model species (e.g. Wu *et al*., 2014; Xu *et al*., 2014; Zhu *et al*., 2014). Comparative transcriptomics allows us to identify potential candidate genes regulating or driving the production of wood in secondarily woody species. This descriptive approach can only provide a 'snap shot' of the situation, but it is a valuable first step in identifying the developmental pathways and changes involved.

The pantropical genus *Begonia* provides an ideal group to investigate the evolution of secondary woodiness. *Begonia* contains at least 1550 species comprising mostly herbaceous species, but also a number of species that grow as woody shrubs (Doorenbos, Sosef & de Wilde, 1998). The derived nature of *Begonia* wood was described by Carlquist (1985) who studied the wood anatomy of four woody South American species and observed characters demonstrating protracted juvenilism, namely the presence of tall multiseriate rays with mainly upright ray cells and wide scalariform intervessel pitting. The estimated number of woody begonias is difficult to assess because of incomplete collection efforts in some regions combined with the lack of thorough regional flora treatments in these areas, but up to ca. 50 woody species might be possible. In published phylogenies, only three woody African, one woody Asian, and six woody neotropical *Begonia* species are included, and these woody species represent at least seven independent shifts towards secondary woodiness, suggesting this is a fairly labile trait within *Begonia* (Plana, 2003; Forrest, Hughes & Hollingsworth, 2005; Goodall-Copestake *et al*., 2010; Thomas *et al*., 2011; Moonlight *et al*., 2015).

Most of the woody *Begonia* species are native to wet tropical mountain peaks of SE Asia (Beaman, Anderson & Beaman, 2001; Hughes & Pullan, 2007), Andean South American (Mark Tebbitt, pers. comm.), moist East African montane forests (Reitsma, 1984; Plana, Sands & Beentje, 2006) or moist tropical West African islands (São Tomé and Principe). Consequently, for a majority of the woody begonias, drought stress is definitely not involved in wood formation, although this has been suggested by recent experimental results based on embolism resistance measures in stems of herbaceous and woody *Arabidopsis thaliana* individuals (Lens *et al*., 2013b). Nevertheless, for some other woody begonias drought stress is an issue, such as some of the South East Asian woody begonias native to dry coralline limestone hills with low water-holding capacity (Kiew, 1998, 2001), and some neotropical woody species inhabiting dry habitats in the Andes. It even appears that the woody Andean species in these drier areas are woodier than the ones growing in more mesic Andean habitats, such as the narrow Peruvian endemic *B. gorgonea* exhibiting strikingly woody rhizomes in xeric environments (Mark Tebbitt, pers. comm.).

The occurrence of secondarily woody species in both wet and dry climates, makes this an excellent model genus to investigate the diverse array of factors that may drive shifts to secondary woodiness (Kiew, 1998, 2001; Beaman *et al*., 2001; Hughes & Pullan, 2007). *Begonia* also has the advantage of having an array of genomic resources including a draft genome sequence, transcriptome datasets, and genetic maps (Brennan

et al., 2012; C. Kidner in prep), making questions about the genetic underpinnings of the evolution and development of secondary woodiness tractable.

Here we take an interdisciplinary approach using three methods to investigate the evolution and development of secondary woodiness in *Begonia*: (1) we describe detailed stem anatomical differences between mainly herbaceous and woody species from Borneo; (2) use an updated phylogenetic analysis focusing on Bornean species to provisionally estimate the number of shifts towards secondary woodiness in Borneo and world-wide; and (3) characterize gene expression in wood forming tissues of the secondarily woody *B. burbidgei* Stapf native to Borneo. These results provide us with preliminary data to inform further discussion on the diversity of environmental and abiotic factors that might be associated with secondary woodiness in *Begonia*.

MATERIAL AND METHODS

Definition of woodiness vs. herbaceousness

We recognize the fuzzy boundary between woodiness and herbaceousness creates difficulties (Dulin & Kirchoff, 2010; Lens *et al*., 2012a, b, 2013a). We are only interested in the evolutionary processes that underlie the dramatic transition from ancestral 'herbaceous' species with no or limited wood formation to derived woody shrubs with extensive wood development. A strict botanical definition of a woody species is lacking, but in practice there are striking differences among groups including 'herbaceous' and 'woody' begonias (Figs 1, 2). We define secondarily woody species as shrubs producing a distinct wood cylinder extending towards the upper stem parts as shown in Figure 4. This criterion only applies to the following species in our sampling: *B. burbidgei*, *B. beryllae* Ridl., *B. fruticosa* A. DC. and two species new to science found in Crocker Range Park (Malaysia; *Begonia* sp. nov. spec. 2 and spec. 3). Based on this definition, all the 45 *Begonia* species that were investigated by Lee's study of stem anatomy in *Begonia* (Lee, 1974) should be called herbaceous, although the author mentioned 'considerable secondary growth' in some species studied. We aim here to investigate the switch to production of a robust wood cylinder extending throughout the stems in some woody species.

Taxonomic sampling, sequencing protocol and phylogenetic analysis

During the September 2012 expedition to Mount Kinabalu and Crocker Range Park, the corresponding author collected material from 14 *Begonia* species including four woody species and three new to science. The frequency of the woody growth habit in this locality suggested this would be a good starting point for investigating the evolution of woody growth within the genus. Voucher material is deposited in the herbaria of Sabah Parks (Sabah, Malaysia) and the Forest Research Institute Malaysia, voucher data are presented in Table S1.

To understand the phylogenetic context of the woody species, the species collected in Borneo were incorporated into a plastid phylogeny of South East Asian begonias (Thomas *et al*., 2011, 2012). Three non-coding plastid DNA regions (*ndh*A intron, *ndh*F–*rpl*32 spacer, *rpl*32–*trn*L spacer), which were shown to be of considerable phylogenetic utility at the inter- and infrasectional level in *Begonia* were amplified (Thomas *et al*., 2011, 2012; Moonlight *et al*., 2015). In total, the ingroup comprised 105 accessions sampled broadly from all major Asian *Begonia* sections. A focus was put on accessions of the large section *Petermannia* (> 250 species), which includes several distinctly woody species. Accessions of the woody *B. beryllae*, *B. burbidgei* Stapf, *B. vaccinioides* and *B. spec.* 2, all of which were derived from our recently collected material, were included. Accessions of seven herbaceous species from Kinabalu, from which 19 species have been described in total (Hughes, 2008), as well as accessions of an additional 42 species from the entire geographic range of section *Petermannia* were included in the analyses. Two African species, *Begonia dregei* Otto and Dietr. and *Begonia sutherlandii* Hook, were selected as outgroup based relationships indicated in previous molecular phylogenetic studies (Plana *et al*., 2004; Goodall-Copestake *et al*., 2010). DNA sequences generated in previous studies (Thomas *et al*., 2011, 2012) were downloaded from the nucleotide database of the National Centre for Biotechnology Information (http://www.ncbi.nlm.nih.gov/), and 55 sequences were newly generated for this study (GenBank accession numbers are listed in Table S2).

Total genomic data was extracted from living material or silica gel dried material using the innuPrep Plant DNA Kit (Analytika Jena, Jena, Germany) according to the manufacturer's protocols. Primers and amplification protocols for the three chloroplast markers were the same as in Thomas *et al*. (2011). Sequencing polymerase chain reaction (PCR) products were purified and sequenced by MACROGEN (Amsterdam) using an AB 3730 DNA Analyser (Applied Biosystems).

Sequences were assembled and edited using *Geneious* v6.1.7 (Drummond *et al*., 2010). The sequences were pre-aligned using the multiple sequence alignment software *MUSCLE* (Edgar, 2004) implemented in *Geneious* using default settings, and subsequently manually checked and optimized in

Figure 1. Overview of variation in herbaceousness (A–C) and distinct woodiness (C, D) among Asian begonias, and growth habits with reference to specialised organs: tuberous (A), rhizomatous (B), non-tuberous/rhizomatous (C–E). A, *Begonia tenuifolia* Dryand, *Begonia obovoidea* Craib (Lens and Tisun 62). C, *Begonia gambutensis* Ardi and DC Thomas D, *Begonia vaccinioides* Sands (photography credit Rogier van Vugt). E, *Begonia* sp. nov. (Lens and Tisun 78).

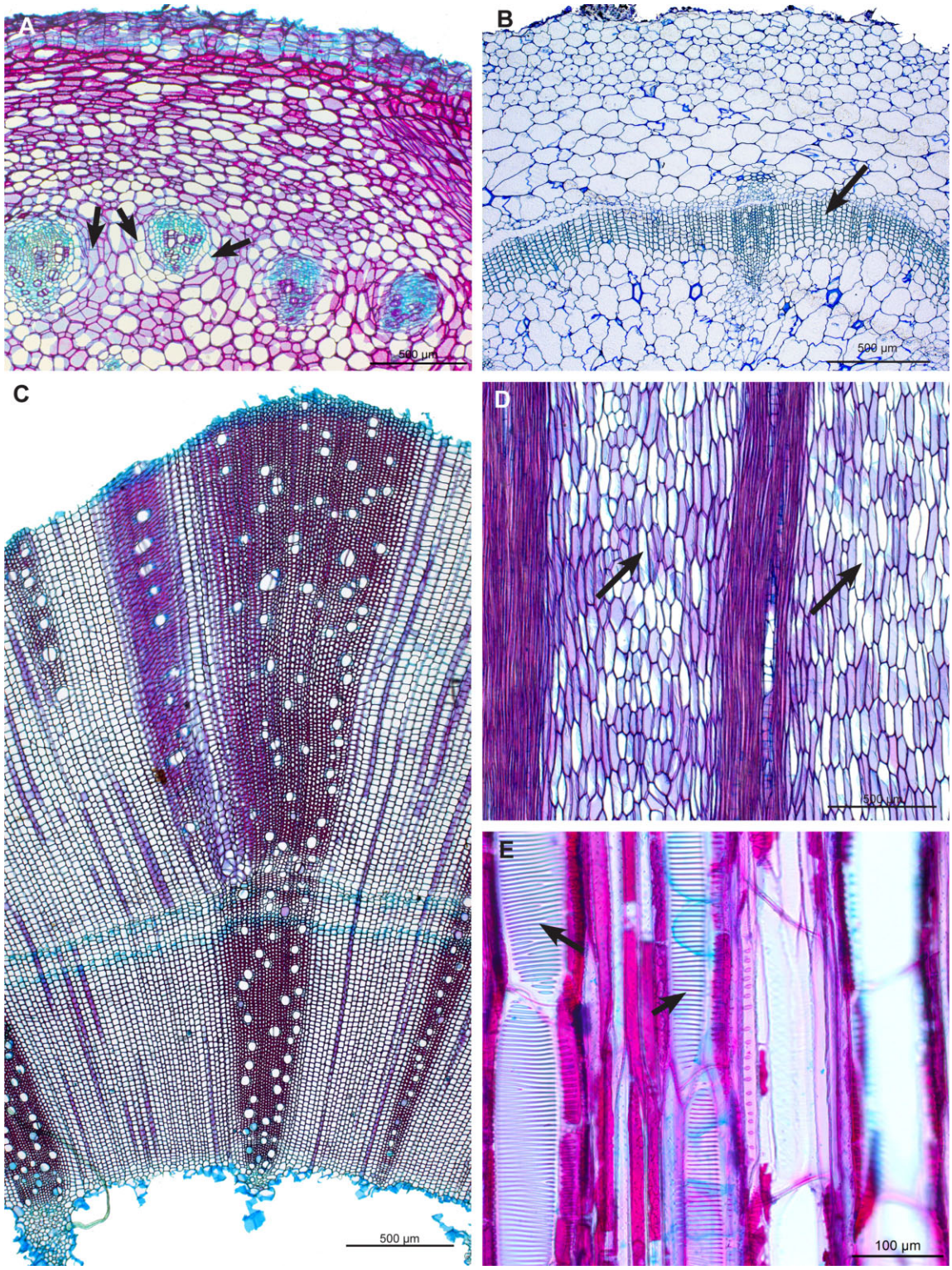

Figure 2. Light microscope sections of *Begonia* stems showing marked anatomical diversity between herbaceous species (A, B) and woody species (C–E). A, *Begonia chlorocarpa*, cross-section showing intact primary vascular bundles, interfascicular cambium is developing (arrows). B, *Begonia* aff. *cauliflora*, cross-section through basal stem part showing narrow wood cylinder (arrow). C, *Begonia* sp. nov. (Lens and Tisun 78), cross-section at the stem base showing marked wood cylinder with tall rays. D, *Begonia* sp. nov. (Lens and Tisun 82), tangential section illustrating tall rays with mainly upright ray cells (arrows). E, *Begonia fruticosa*, tangential section showing wide gaping scalariform intervessel pitting (arrows).

Geneious. Inversions were identified in the *ndh*F-*rpl*32 spacer region of all Philippine samples of *Begonia* section *Diploclinium* (309–355 bp), the *rpl*32-*trn*L spacer of *Begonia pendula* Ridl. (37 bp), as well as in the *rpl*32-*trn*L spacer of six distantly related species (11 bp, see Thomas *et al*., 2011). These inversions were reverse-complemented, thereby retaining substitution information in the fragments.

Bayesian phylogenetic reconstructions were performed using the XSEDE application of *MrBayes* v3.2.2 (Ronquist & Huelsenbeck, 2003) provided by the CIPRES Science Gateway (Miller, Pfeiffer & Schwartz, 2010). Three partitions based on spacer and intron identity (*ndh*A intron, *ndh*F-*rpl*32 spacer, *rpl*32-*trn*L spacer) were defined *a priori*. Models of sequence evolution of each nucleotide sequence partition were determined using *MrModelTest* (Nylander, 2004) under the Akaike Information Criterion (AIC). Parameters for character state frequencies, substitution rates of nucleotide substitution models, and rate variation among sites were unlinked across partitions. The mean branch length prior was set from the default mean (0.1) to 0.01 to reduce the likelihood of stochastic entrapment in local tree length optima (Brown *et al*., 2010; Marshall, 2010). Four independent Metropolis-coupled MCMC analyses were run. Each search of 10 million generations used four chains, a temperature parameter setting of 0.6 and was sampled every 1000 generations. Convergence was assessed by using the standard deviation of split frequencies with values < 0.005 interpreted as indicating good convergence. *Tracer* v1.5 (Rambaut & Drummond, 2009) was used to check for stationary and adequate effective sample sizes for each parameter (ESS > 200). Convergence of posterior probabilities of splits from different runs were checked using the Compare and Cumulative functions of *AWTY* (Nylander *et al*., 2008). The initial 25% of samples of each run were discarded as burnin and the remaining trees were summarized as 50% majority-rule consensus tree with nodal support summarized as posterior probabilities.

Stem anatomy

Stem samples of the four woody species collected on Mount Kinabalu (*B. beryllae, B. burbidgei, B. sp2* and *B. sp3*) were compared to stem samples from herbaceous species in the same area and sampled more widely across SE Asia. A stem sample from a fifth woody species, the large liana *B. fruticosa* (Sao Paolo, Brazil), was included to address the range of wood anatomy in *Begonia*. In total, eight wood samples from the five species investigated were sectioned using a sliding microtome. Wood sections were coloured with safranin-alcian blue mix and mounted with euparal (standardized protocol explained in Lens *et al*., 2007). Nine samples representing the basal stem part of herbaceous species were embedded in LR White resin (hard grade, London Resin, UK), sectioned with a rotary microscope, and stained with toluidine blue according to the protocol described in Hamann, Smets & Lens (2011). Transverse sections and longitudinal sections were made for the woody species, while the herbaceous species were represented by transverse sections only.

Stem transcriptome analysis

Samples from a woody *B. burbidgei* individual were collected on the Mount Kinabalu summit trail at 2870 m asl in ultramaphic soils (Lens and Tisun 51, Fig. 3, see Table S1 for detailed voucher data). Green and more basal woody stem samples were collected (Fig. 4A), sliced longitudinally into smaller fragments, and stored immediately in RNA *later* (Ambion). Samples from the same regions of the same stems were also collected for histological analysis. Figure 4B shows the green tissue sample with cambial growth just initiating and Figure 4C shows much more wood formation in the woody stem sample. Material used for transcriptome sequencing was from the same stem section as for the histological analysis.

Due to the difficulty in determining the cambial layer in the preserved tissue sections, material from the entire stem cylinder (as shown in Fig. 4a, c) was used for the RNA preps. RNA was extracted using Invitrogen's Plant RNA extraction solution with modifications to the protocol to include an acid phenol extraction after the chloroform extraction and a final LiCl precipitation. Illumina-compatible sequencing libraries were made using the Illumina Truseq RNA Sample Preparation Kit according to the manufacturer's protocol. Samples were multiplexed 6 per lane of HighSeq 2000. 161 874 378 raw reads were generated from the two samples. Reads were quality trimmed using FASTQ groomer in Galaxy (cut-off value 20, minimum percentage 90; Blankenberg *et al*., 2010) and the cambial stage (28 221 304 reads) and woody stage (49 756 127 reads) reads were jointly assembled into 98 258 contigs using Trinity (Haas *et al*., 2013). 24 707 contigs were over 1 kb, 1525 over 3 kb. 280 chloroplast and mitochondrial sequences were removed by comparison to the *Begonia* chloroplast genomes and previously identified mitochrondrial sequences (Brennan *et al*., 2012; Harrison, 2012). Reads were mapped back to the assembly using bowtie 2.0 (Langmead & Salzberg, 2012). We used blast matches to the draft *Begonia conchifolia* A. Dietr. genome (C. Kidner, unpublished) to the TAIR set of *Arabidopsis thaliana* proteins

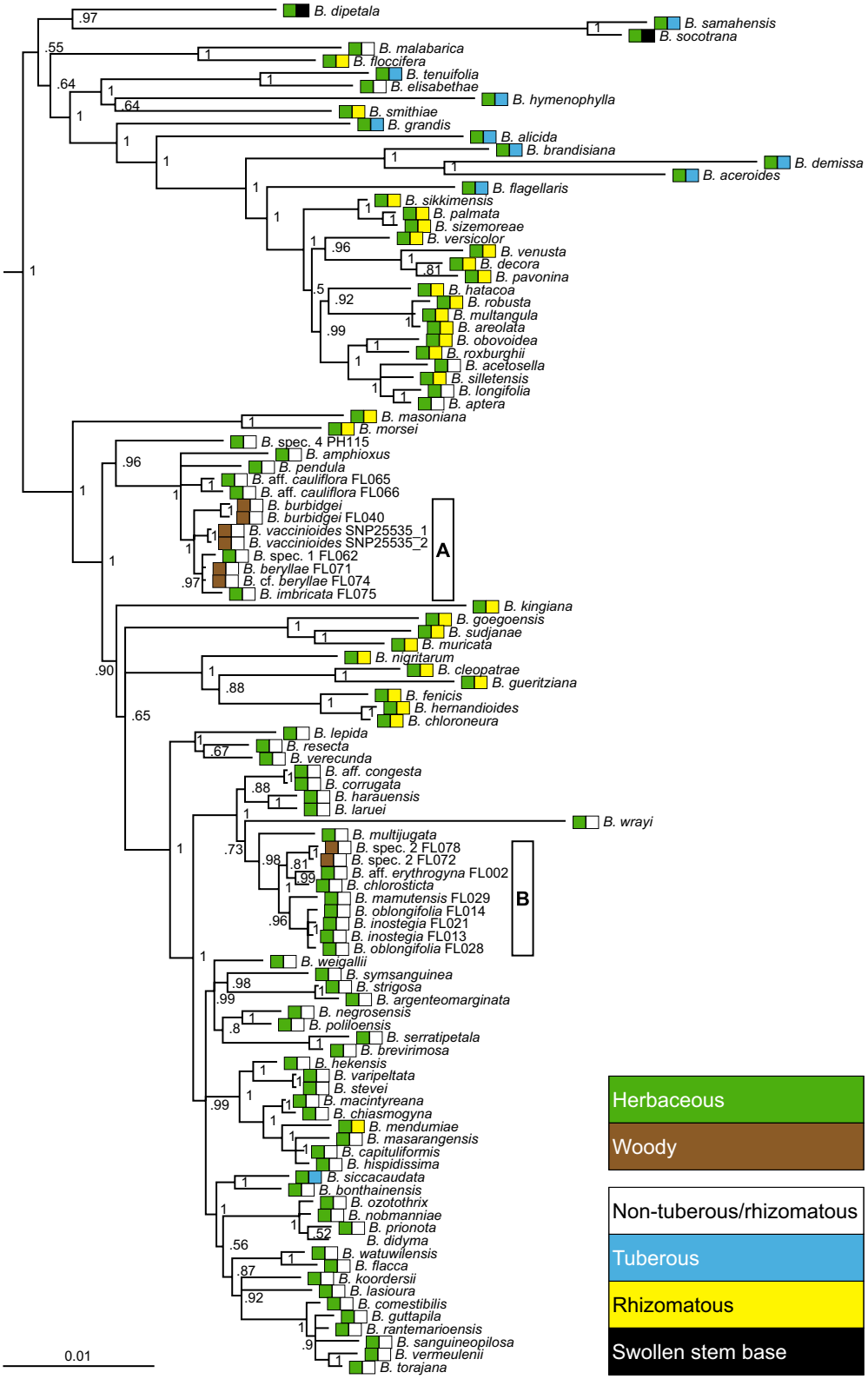

Figure 3. Bayesian 50% majority-rule consensus tree (cpDNA data: *ndh*A intron, *ndh*F-*rpl*32, *rpl*32-*trn*L; three data partitions; 110 accessions). Posterior clade probabilities are indicated next to the nodes. Squares next to the terminals indicate growth habit types: black, swollen stem base; blue, tuberous; brown, woody; green, herbaceous; purple, rhizomatous; white, non-tuberous/–rhizomatous.

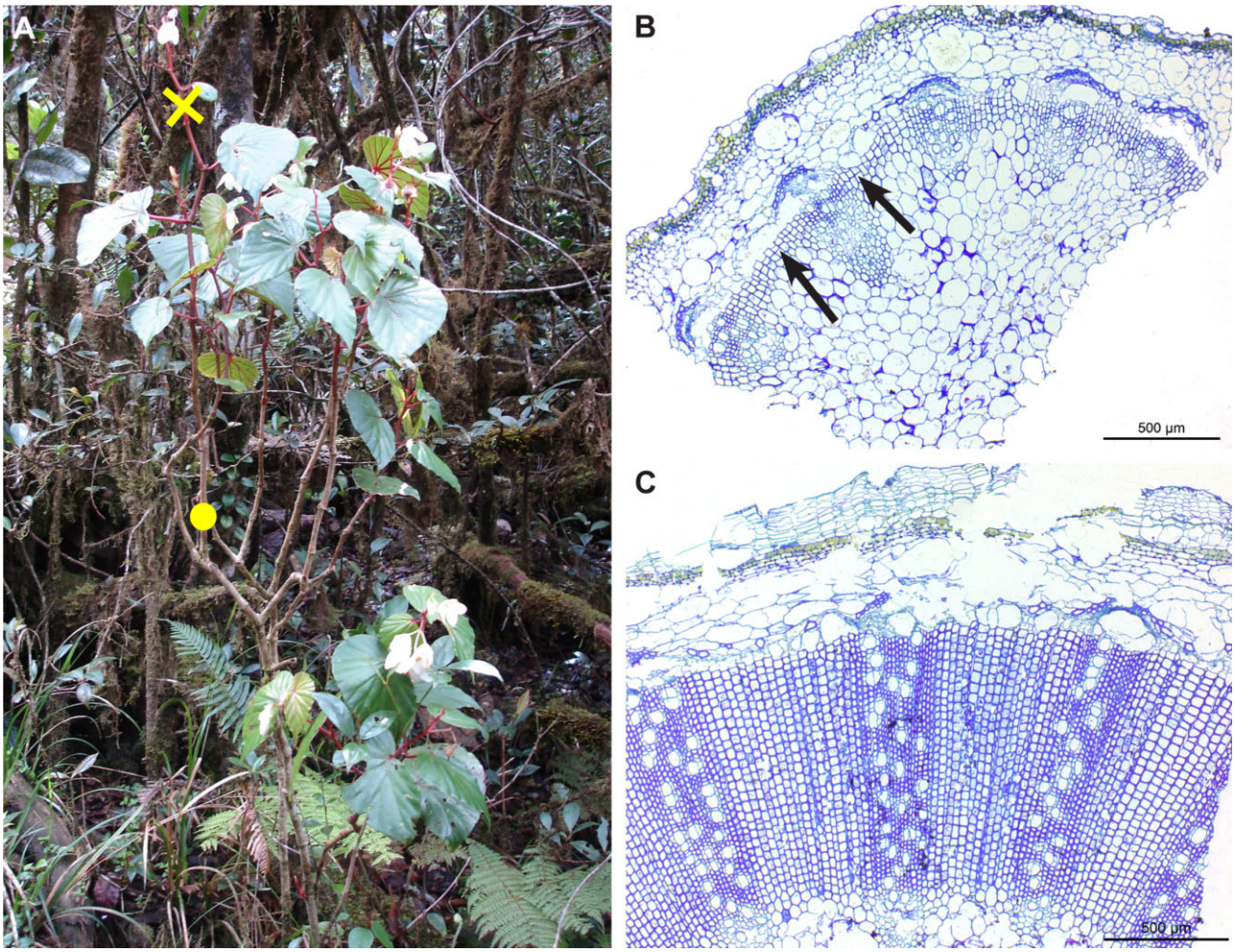

Figure 4. Woody *Begonia burbidgei* individual that was used for stem transcriptome analysis (A) and its stem developmental stages (B, C) sampled in RNA later in the field (mount Kinabalu summit trail, Sabah, Borneo). A, Habit (Lens and Tisun 51). B, Cross-section through inflorescence stem that has initiated a vascular cambium (arrows, yellow cross in A) – cambial stage transcriptome. C, Cross-section through more basal stem part showing extensive wood formation (yellow circle in A) – wood stage transcriptome.

(TAIR10_pep_20101214_updated at http://www.arabidopsis.org) and the *Populus trichocarpus* Torr. and A. Gray protein dataset from phytozome (http://www.phytozome.com, Ptrichocarpa_210_protein.fa.gz) to annotate the sequences. 54 030 sequences had a blastn hit at 1e-40 or better in the *Begonia conchifolia* draft genome to 17 743 unique ORFs. As *B. conchifolia* is a distant relative of *B. burbidgei* (Thomas *et al.*, 2012), we expect this set to include conserved *Begonia* genes. 25 042 sequences had a blastx hit at 1e-10 or better to 10 423 unique TAIR proteins. The TAIR hits were used to annotate the *B. burbidgei* sequences. RSEM and edge-R were used to detect genes differentially expressed between the cambium stage and wood stage samples (Robinson, McCarthy & Smyth, 2010; Li & Dewey, 2011; Haas *et al.*, 2013).

We compared coding sequences for orthologous genes in *B. burbidgei* and the herbaceous *B. venustra*, for which a transcriptome dataset from vegetative buds has already been published (Brennan *et al.*, 2012) using custom scripts produced by K. Emelianova (available on request). To run Reciprocal Best BLAST Hits to identify orthologs in the two species, longest open reading frames were extracted from ortholog pairs using getORF (Rice, Longden & Bleasby, 2000) and translationally aligned using MAFFT (Katoh & Standley, 2013), Virtual Ribosome (Wernersson, 2006) and RevTrans (Wernersson & Pedersen, 2003). Translational alignments were used as input to the Codeml package of PAML (Yang, 2007). The site test was used in Codeml to test for positive selection, using the one-ratio model, producing values of kN, kS and kN/kS for each orthologous

pair. Orthologs which were more than 75% divergent and thus unsuitable for the NG86 method were not included in the final results. The ortholog pairs with indications of positive selection were annotated by blastx to *Arabidopsis* proteins; only those with a match of at least 1e-10 or better were included in the rest of the analysis.

We used reciprocal tblastx to identify orthologs between the *Begonia burbidgei* and *Populus trichocarpus* sequences from a published cambium transcriptome (Liu *et al*., 2014), and compared relative expression levels by reads per million reads per kb of the *Populus* coding sequence (as *Begonia burbidgei* sequences were often not the full coding sequence) (Table S6).

RESULTS

Variation in habit

Secondarily woody *Begonia* species are characterized by a shrubby growth habit and can usually be easily distinguished from herbaceous relatives (Figs 1, 3). In general, herbaceous species have semi-succulent stems, and may have specialized storage organs such as tubers (Fig. 1A) or rhizomes (Fig. 1B) (Figs 1C, 3). In contrast, secondarily woody species are shrubs (Fig. 1D, E) characterized by wood formation extending towards the upper parts of the stem; rhizomes or tubers are generally absent in the Bornean clade we examined, though they seem to occur in some neotropical woody *Begonia* (Carlquist, 1985; Fig. 3; see also Materials and Methods for further discussion about the definition between woodiness vs. herbaceousness).

Description of wood anatomy in *Begonia*

In order to develop a clade of Bornean *Begonia* as a model for understanding the secondary evolution of woodiness we investigated mature wood anatomy in detail for four of these woody species from samples collected in the field (*B. burbidgei* Stapf (Lens and Tisun 44), *B. beryllae* Ridl. (Lens and Tisun 71), *B.* sp. nov. 2 and 3 (Lens and Tisun 78 and 82). The only pre-existing description of wood anatomy in *Begonia* is for three neotropical species and a hybrid (Carlquist, 1985). We included the large neotropical liana *B. fruticosa* A. DC. to allow direct comparisons of our South East Asian species with an unrelated woody species from a parallel radiation (Fig. 2, Table S1). The wood description for all these samples is similar and can be summarized as follows:

Growth ring boundaries absent. Wood diffuse porous. Vessels concentrated in the intrafascicular regions, (7)-17-59-(82)/mm^2, usually solitary, sometimes in short radial multiples of 2–3 and occasionally in short tangential multiples of 2-(3), vessel outline angular. Vessel perforation plates simple. Lateral wall pitting typically wide gaping scalariform (Fig. 2E), pits with minute borders, pit cavities 18–50 μm in horizontal size, non-vestured. Tangential vessel diameter (20)-30-100-(130). Vessel elements (150)-220-380-(470) μm long. Length-on-age curve of vessel elements flat or slightly decreasing. Tyloses present in *B. fruticosa* and occasionally in *B.* sp. nov. 3. Tracheids absent. Fibres sometimes septate, occasionally septate in *B. fruticosa*, thin-walled, (250)-430-520-(680) μm long, with mostly simple to occasionally minutely bordered pits distributed in radial and tangential walls. Axial parenchyma scanty paratracheal, 2-4-(5) cells per strand. Rays exclusively multiseriate and confined to the interfascicular regions, (16)-20-30-(36) cells wide, and very tall (Fig. 2C, D), at least 1500 μm but often much higher than the length of the sections, 0–2 rays/mm^2. Exclusively upright ray cells present in *B. burbidgei*, *B. beryllae* and *B.* sp. nov. 2 and 3 (Fig. 2D), but mainly procumbent to sometimes also square to upright in *B. fruticosa*. Sometimes thick-walled sclereids in the tall rays. Sheath cells and mineral inclusions not observed. A tendency towards layering in fibres, vessels elements and axial parenchyma strands in *B. fruticosa*.

Molecular phylogenetics

Three non-coding plastid DNA regions (*ndh*A intron, *ndh*F–*rpl*32 spacer, *rpl*32–*trn*L spacer) from 105 Asian species and two African species were used for a Bayesian phylogenetic reconstruction of the sampled Bornean *Begonia* species (Table S2; Fig. 3). Secondarily woody species native to Borneo, all of which are assigned to *Begonia* section *Petermannia*, are retrieved in two distantly related clades of Bornean begonias: clades A and B. The strongly supported clade A (posterior probability, PP: 1) includes accessions of the woody species *B. burbidgei* and *B. vaccinioides* Sands, the woody species *B. beryllae* Ridl. and accessions of the herbaceous species *B. imbricata* Sands and *B.* spec. 1. Clade B (PP: 0.98) includes the woody species *B.* spec. 2, as well as several herbaceous species (*B. chlorosticta* Sands, *B.* aff. *erythrogyna* Sands, *B. mamutensis* Sands, *B. oblongifolia* Stapf, *B. inostegia* Stapf).

Stem transcriptomics of *Begonia burbidgei*

Tissue from young and older stages of a stem of one *B. burbidgei* individual was collected from a mature Mount Kinabalu individual (Fig 4A, Lens and Tisun 5, see Table S2 for more details on individual and locality), and total RNA was extracted. Histology of the samples showed that the younger sample had just initiated cambial growth (shown in Fig. 4B) and the older sample had extensive woody growth (Fig. 4C). Transcriptomes were produced from each sample (all tissues in a cross-section of the stem) and annotated

by comparison to *Arabidopsis thaliana* Heynh proteins and *Populus trichocarpa* Torr and A. Grey.

We used three search strategies to mine this data for information on the molecular wood development in this species. The first approach was to use RSEM and edge-R (Robinson *et al*., 2010; Li & Dewey, 2011; Haas *et al*., 2013) to detect genes differentially expressed between the initiating and fully woody stems of *B. burbidgei*. These genes are expected to include those promoting vascular cambium activities versus wood differentiation in *Begonia*. The second approach was to compare coding sequences for orthologous genes in *B. burbidgei* and the herbaceous *B. venustra*, for which a transcriptome dataset from vegetative buds has already been published (Brennan *et al*., 2012). Genes showing a high Ks/Kn ratio in this comparison are expected to include the targets of selection during the evolution of woody growth. Our third approach was to compare the genes expressed in the primarily woody *Populus trichocarpa* cambium with those found in secondarily woody *B. burbidgei* stems. Genes found in *P. trichocarpa* but missing from the *B. burbidgei* sample are expected to include those which are involved in cambium activity in the primarily woody *P. trichocarpa* but not in the secondarily woody *B. burbidgei*. We also examined the expression levels in *B. burbidgei* of genes which had been implicated in secondary growth and lignin biosynthesis in a range of model species.

Comparative transcriptomics of wood and cambium stage *B. burbidgei* stems

Edge-R analysis produced a list of 577 contigs showing differential expression between wood and cambium stage stems by more than two-fold and a *P*-value of less than 1e-3. These contigs were annotated by blastx matches to the *Arabidopsis thaliana* and the *P. trichocarpa* protein databases (Table S1). 287 of these 577 contigs had no good blastx match (at 1e-10 or better) in *Arabidopsis* or poplar. These sequences will require further analysis to determine function.

The genes with best support for differential expression between the wood and cambium stage of *Begonia burbidgei* stems and with good annotation include a number involved in meristem determinacy (REBELOTE AT3G55510.1), growth (the expansin AT1G26770.1), cell wall biosynthesis (xyloglucan endotransglycosylase AT4G25820.1), many lipid metabolism associated genes including HOTHEAD (AT4G25820.1) and FIDDLEHEAD (AT2G26250.1) as well as some genes thought to be involved in secondary cell wall synthesis such as the FASCICLIN-like arabinogalactan protein 8 (FLA8) (AT2G45470.1) (Table S3).

The Gene Ontology (GO) terms of the differentially expressed annotated contigs were counted and compared to the annotations for all *B. burbidgei* sequences using agriGO (Du *et al*., 2010). 26 GO terms were enriched in the differentially expressed set (Table 1). These include carbohydrate transporters, cell wall associated genes and serine-type exopeptidases.

Comparative transcriptomics of the woody *B. burbidgei* and the herbaceous *B. venustra*

B. venustra King is an herbaceous species from Malaysia, not closely related to the woody *B. burbidgei* (Thomas *et al*., 2012). A transcriptome for vegetative buds of *B. venustra* had been previously analysed (Brennan *et al*., 2012). The vegetative bud sample includes some young stem tissue but the range of cell types present differ too much from the B. *burbidgei* stem samples for differential expression analysis to be useful. However, comparison of the sequences suggests some interesting variation between these two species.

A python pipeline was used to analyse sequence variation between orthologous genes from the woody *B. burbidgei* and the herbaceous *B. venustra*. The 256 orthologs showing kN/kS ratios exceeding 1 – suggestive of positive selection – are listed in Table S4. This set includes interesting candidate regulators of woody growth such as orthologs of two myb transcription factors (AT5G41020 and AT1G26780 (LOF1)). It also includes orthologs of genes that may be linked to growth on ultramafic soil such as AT4G34050, an enzyme involved in response to cadmium, and orthologs of secondary metabolism enzymes such as caffeoyl coenzyme A O-methyltransferase 1 (AT4G34050) (Table S4). The GO terms for those with a Kn/Ks over 1 were compared to the annotations for all pairs using agriGO (Du *et al*., 2010). No significantly over-represented term was identified. We find no evidence supporting the hypothesis that the evolution of woody growth in *B. burbidgei* involved concerted protein sequence change for genes in a particular pathway.

Comparative transcriptomics of *B. burbidgei* and *Populus trichocarpa*

To compare wood development in *B. burbidgei* to that in model wood forming species, we used the genetic resources of *P. trichocarpa* (Tuskan *et al*., 2006). 3974 *B. burbidgei* contigs had a blastx hit to a *P. trichocarpa* protein but did not have matches (at 1e-40) in the *B. conchifolia* genome. This set of potential wood-associated genes comprised matches to 2522 *P. trichocarpa* genes including meristem regulatory

Table 1. Gene Ontology (GO) terms enriched in the set of genes differentially expressed between cambium stage and wood samples of *B. burbidgea* stems

GO term	Description	Count in significantly differentially expressed list	Count in *B. burbidgei* transcriptome	*P*-value	FDR
GO:0015295	Solute:hydrogen symporter activity	13	70	6.90E–11	6.90E–09
GO:0005402	Cation:sugar symporter activity	13	70	6.90E–11	6.90E–09
GO:0005351	Sugar:hydrogen symporter activity	13	70	6.90E–11	6.90E–09
GO:0051119	Sugar transmembrane transporter activity	13	80	4.00E–10	2.90E–08
GO:0015144	Carbohydrate transmembrane transporter activity	13	85	8.60E–10	5.10E–08
GO:0015293	Symporter activity	13	91	2.00E–09	8.70E–08
GO:0015294	Solute:cation symporter activity	13	90	1.80E–09	8.70E–08
GO:0015291	Secondary active transmembrane transporter activity	14	188	1.90E–06	7.20E–05
GO:0008236	Serine-type peptidase activity	9	88	9.80E–06	0.00029
GO:0017171	Serine hydrolase activity	9	88	9.80E–06	0.00029
GO:0022892	Substrate-specific transporter activity	25	624	3.00E–05	0.0008
GO:0022891	Substrate-specific transmembrane transporter activity	22	534	5.50E–05	0.0014
GO:0015075	Ion transmembrane transporter activity	18	392	6.20E–05	0.0014
GO:0008324	Cation transmembrane transporter activity	15	293	7.20E–05	0.0015
GO:0070008	Serine-type exopeptidase activity	5	30	8.90E–05	0.0016
GO:0008238	Exopeptidase activity	6	48	9.50E–05	0.0016
GO:0022857	Transmembrane transporter activity	25	672	9.30E–05	0.0016
GO:0004185	Serine-type carboxypeptidase activity	5	30	8.90E–05	0.0016
GO:0004180	Carboxypeptidase activity	5	32	0.00012	0.0019
GO:0022804	Active transmembrane transporter activity	17	399	0.00023	0.0033
GO:0005215	Transporter activity	29	880	0.00022	0.0033
GO:0070011	Peptidase activity, acting on L-amino acid peptides	11	266	0.0032	0.044
GO:0005576	Extracellular region	11	97	3.80E–07	7.80E–05
GO:0048046	Apoplast	8	63	6.00E–06	0.00062
GO:0030312	External encapsulating structure	12	213	0.00015	0.0077
GO:0005618	Cell wall	12	209	0.00012	0.0077

genes such as a ZPR2 ortholog (Potri.002G149600) and arabinogalactans (Table S5).

We used expression data from *P. trichocarpa* cambium (Liu *et al*., 2014) to find genes expressed during wood development in *P. trichocarpa* that are differentially expressed or missing in the wood of *B. burbidgei* (Table S6). A number of interesting genes were differentially expressed in this comparison. Potri.014G015600, an ortholog of the myb transcription factor NtLIM, required for full activation of the phenylpropanoid pathway (Kawaoka & Ebinuma, 2001), is expressed at over 1000-fold higher in *P. trichocarpa* cambium than in *B. burbidgei* stems.

Agusti *et al*. (2011) identified MOL1 and RUL1 as opposing regulators of secondary growth. These two genes are expressed at similar levels in the cambium stage stem and the woody stem of *B. burbidgei* and the *P. trichocarpa* cambium sample, but orthologs of three of the targets identified in Agusti *et al*. (2011) – At1g46480 (WOX1), At1g52340 (ABA2) and At5g57130 – are upregulated in the *Begonia* samples, suggesting this pathway may also be modified in *B. burbidgei* wood development.

This analysis also highlighted a set of genes expressed in the *P. trichocarpa* cambium sample but not recovered from the *B. burbidgei* transcriptomes. Most of these genes are not annotated in *P. trichocarpa* but some suggest interesting differences between *Populus* and *Begonia* wood. Some arabinogalactans, in particular FASCICLIN-like

Table 2. The reads per kilobase per million reads (RPKM) for genes encoding enzymes in the lignin biosynthetic pathway for *Begonia burbidgei* (wood and cambium stage samples) compared to a sample of *Populus trichocarpus* cambium

Enzyme	Direction	Enzyme code	Wood stage	Cambium stage	Poplar cambium
Phenylalanine ammonia-lyase	To lignin	4.3.1.24	333.6	281.76	474.6
Trans-cinnamate 4-monooxygenase	To lignin	1.14.13.11	522.97	388.13	1466.45
4-Coumarate-CoA ligase	To lignin	6.2.1.12	175.85	118.93	2.44
Caffeoyl shikimate esterase	To lignin	3.1.1.5	59.69	63.27	60.1
Shikimate *O*-hydroxycinnamoyltransferase	To lignin	2.3.1.133	95.01	73.58	438.17
Caffeoyl-CoA *O*-methyltransferase	To lignin	2.1.1.104	11.63	9.35	1857.5
Cinnamyl-alcohol dehydrogenase	To lignin	1.1.1.195	3.78	5.61	83.9
Peroxidase	To lignin	1.11.1.7	645.27	484.26	1130.42
4-Coumarate-CoA ligase	Away from lignin	6.2.1.12	175.85	118.93	2.44
Beta-glucosidase	Away from lignin	3.2.1.21	114.89	147.88	8.34
Coniferyl-aldehyde dehydrogenase	Away from lignin	1.2.1.68	43.88	45.26	16.63

arabinogalactan proteins were not recovered from the *Begonia* samples though they were expressed at a high level in *Populus* cambium, as were a number of calmodulin-like proteins (Table S7).

To provide an overview of the biosynthetic pathways that differed between the *B. burbidgei* and *P. trichocarpa* samples, we summed the reads per kilobase per million reads (RPKMs) for each enzyme class (EC); these data are presented in Table S8. A key enzyme in the lignin biosynthetic pathway, Caffeoyl-CoA O-methyltransferase, is one of the enzymes much more highly expressed in *Populus* cambium than in the wood or cambium *Begonia* stem samples, whereas 4-coumarate-CoA ligase, one of the earlier enzymes in the pathway is expressed at much higher levels in the *B. burbidgei* samples.

To focus on the lignin pathway, we compared expression levels of all genes from this biosynthetic pathway we could recover from *B. burbidgei*, including the newly identified lignin biosynthetic gene Caffeoyl shikimate esterase (Vanholme *et al*., 2013). The RPKM for genes encoding enzymes in the lignin biosynthetic pathway are listed in Table 2. *B. burbidgei* wood expresses many of the lignin biosynthetic enzymes, but at a lower level than in the *P. trichocarpa* cambium sample, particularly at the distal end of the pathway, whereas the expression level of enzymes leading away from lignin is higher.

Some genes have been proposed as regulators of wood synthesis (reviewed in Andersson-Gunnerås *et al*., 2006). Of the 22 genes listed in Andersson-Gunnerås *et al*. (2006), including many myb transcription factors and some cellulose synthase genes, *B. burbidgei* orthologs were identified for 17 but none of these showed any significant difference in expression levels between our cambium and wood samples (data not shown).

DISCUSSION

High number of convergent shifts towards secondary woodiness within *Begonia*

Our phylogenetic analysis is in agreement with other available molecular phylogenies showing that all woody *Begonia* species have been derived from herbaceous relatives (Plana, 2003; Forrest *et al*., 2005; Goodall-Copestake *et al*., 2010; Thomas *et al*., 2011, 2012). The derived origins of woodiness in *Begonia* species is also supported by the wood anatomy of the five species studied in this paper and the four neotropical *Begonia* shrubs investigated by Carlquist (1985). All these wood samples show protracted juvenilism, as demonstrated by the tall rays with mainly upright ray cells, wide gaping scalariform intervessel pitting, and the flat length-on-age curve for vessel elements (Fig. 2; Carlquist, 1985, 2009, 2012). However, protracted juvenilism in wood is not always associated with secondary woodiness, as some studies have indicated a strong link between protracted juvenilism in wood and specific growth form types in primarily as well as secondarily woody angiosperms (Carlquist, 2009; Dulin & Kirchoff, 2010; Lens *et al*., 2013a).

The number of shifts towards secondary woodiness in *Begonia* is still unknown. The densest phylogenetic sampling of South East Asian begonias (in total ca. 650 species) published to date includes only one woody species, *B. burbidgei* from Borneo (Thomas *et al*., 2012). Our analysis adds the woody *B. vaccinioides*, *B. beryllae*, an unknown woody species collected in the Crocker Range (*Begonia* sp. 2, Lens and Tisun 78), and additional accessions of *B. burbidgei* (Fig. 3), together with original sequences from nine herbaceous Bornean species. Based on our increased sampling, *B. burbidgei* falls together with

B. vaccinioides and *B. beryllae* in the same Bornean clade, while *Begonia* sp. 2 is placed in an unrelated Bornean clade. In the first clade, we hypothesize that these three woody species represent one shift towards secondary woodiness, although *B. beryllae* seems to be more closely related to the herbaceous *B. imbricata* and an unknown herbaceous species in the Crocker Range (Lens and Tisun 62) than to *B. vaccinioides* and *B. burbidgei*. Further increasing the Bornean sampling – more than 100 Bornean *Begonia* species are currently accepted (Hughes, 2008; Sang, Kiew & Geri, 2013) – will identify whether the two herbaceous species are secondarily herbaceous or whether this first Bornean clade includes more than one shift towards secondary woodiness. In the second Bornean clade, the woody *Begonia* sp. nov. 2 clusters together with herbaceous species, with *B. erythrogyna* Sands and *B. chlorosticta* as closest relatives (Fig. 3). It is likely that more shifts in SE Asia have occurred, because there are several other potentially woody Asian species such as *B. keithii* Kiew (Borneo) and *B. merrittii* Merr. (Philippines), for which DNA sequence data are not available to date. *B. merrittii*, native to moist montane forests in the Philippines (Hughes & Pullan, 2007), is definitely woody, but for *B. keithii* and several other species anatomical observations are required to assess whether these species represent truly woody shrubs, or whether wood formation is only limited to the base of the stem as is the case for many herbaceous species (see definition of woodiness vs herbaceousness in Methods section).

Besides the two shifts found in the SE Asian clade, there are also two clear shifts towards secondary woodiness in the African clade (Plana, 2003), the continent on which *Begonia* initially diversified (ca. 160 sp.). One shift leads to the tall shrubs *B. baccata* Hook and *B. crateris* Exell native to the tropical West African islands São Tomé and Principe (Reitsma, 1984), and the second shift is represented by the woody climber *B. meyeri-johannis* Engl. inhabiting moist montane East African forests (Plana *et al*., 2006).

In South America, the situation is more complex due to the high number of species (ca. 690 sp.), the abundance of woody species (possibly up to ca. 30 sp.; Mark Tebbitt, pers. comm.), the lack of taxonomic revisions or detailed stem anatomical observations, and the variation in habit ranging from herbaceous species, towards different types of woody growth forms, such as acaulescent suffrutescent growth forms with thick woody rhizomes (e.g. *B. gorgonea* Tebbitt), woody shrub-like species (e.g. *B. parviflora* Schott), and even tall woody lianas reaching the canopy (*B. fruticosa*). In the recent phylogenetic analysis of neotropical begonias (Moonlight *et al*., 2015), six woody species are included, leading to at least four additional shifts towards secondary woodiness. Again, this number is likely an underestimation and a much denser sampling together with a detailed anatomical survey is desired to obtain a more realistic view on the plasticity of growth forms within *Begonia*.

Towards unravelling the regulation of wood formation in *Begonia*

After the *Helianthus* (Asteraceae) transcriptome dataset of seedlings (Moyers & Rieseberg, 2013), our *Begonia burbidgei* transcriptome dataset is the second study that performs an RNA-seq experiment in stems of a secondarily woody species. While our sampling is limited, our work serves as an initial investigation into the genetic changes which accompany the development of wood in this species.

The induction of genes associated with secondary growth, such as expansins and xyloglucans, is to be expected during the establishment of wood formation in *B. burbidgei* and we do see such changes (Table S3). The differential expression of lipid metabolism genes is more surprising, though some trees produce wood with substantial amounts of lipids (Hoch, Richter & Körner, 2003) and lipid vesicles are involved in transport of lignin precursors to the cell wall. Alternatively, the differences in lipid-related gene expression may be due to changes in tissues other than the cambium. As the stem matures, changes would also be expected in the periderm (Fig. 4C) and this suberized tissue is a likely location for products of the lipid metabolic pathway.

One aim of our work was to identify potential candidate genes for the control and development of secondary growth in *Begonia burbidgei*. Due to sampling constraints we cannot establish strong correlations between particular regulators and wood development, but by using a variety of approaches we have a short list of six sets of genes worthy of further investigation. Firstly, an ortholog of *REBELOTE* is upregulated in wood stages of the *B. burbidgei* stem (Table S3). This gene promotes meristem determinacy in *Arabidopsis* redundantly with *ULTRAPETALA* and *SQUINT* (Prunet *et al*., 2008). In *Begonia conchifolia* six loci encode *REBELOTE*-like genes, only one of which has any transcripts in the vegetative bud transcriptome, and that at a very low level. Based on alignments of the coding regions, this paralog expressed in the vegetative bud is likely the ortholog of the *REBELOTE* gene expressed at high levels in the wood sample of *B. burbidgei*. This ortholog is unlikely to have the same role in terminating meristems as its *Arabidopsis* ortholog as the vegetative buds it is expressed in are active. A change in role could be related to the lack of *ULTRAPETALA* and *SQUINT* co-expression in *Begonia*. None of the

37 *Arabidopsis* genes co-expressed with *REBELOTE* according to GeneMANIA (genemania.org) have *B. burbidgei* orthologs with significantly differential expression between the two samples ($P < 0.05$), suggesting the *B. burbidgei* ortholog of *REBELOTE* is active in a different pathway to the *Arabidopsis* original.

A second gene that could be involved in the maintenance of an active cambium is the ortholog of *ZPR2* found in *B. burbidgei* and *P. trichocarpa* but not in the genome of the herbaceous *B. conchifolia* (Table S5). *ZPR2* is part of a gene family, which interacts with HD-ZIPIII proteins to regulate meristem function in *Arabidopsis* (Wenkel *et al*., 2007). Association of expression variation in these genes and woody growth would establish whether these regulators are key points in the evolution of woody growth in *Begonia*.

Another of the differentially expressed genes in the woody *B. burbidgei* without orthologs in the genome of the herbaceous *B. conchifolia* is an ortholog of AT5G23960, *TERPENE SYNTHASE 21* (TPS21). This gene encodes a sesquiterpene synthase involved in generating all of the group A sesquiterpenes found in the *Arabidopsis* floral volatile blend. Its *B. burbidgei* ortholog (comp104430_c1_seq1) is expressed at very high levels in the vascular cambium stem stage, but goes down in the woody stage. The *Begonia* ortholog could be involved in generating secondary products associated with woody tissue.

Fourthly, orthologs of two myb transcription factors (AT5G41020 and AT1G26780 (LOF1)) show signs of sequence divergence between the herbaceous *B. venustra* and woody *B. burbidgei* (Table S4). Such divergence could change their targets or their co-regulators and so completely change their effects. Myb transcription factors that regulate wood development have already been characterized in a number of species (Goicoechea *et al*., 2005; Zhang *et al*., 2014). Further analysis of this pair through transgenic experiments could show if they have the capacity to affect wood development.

Another potential regulator of *B. burbidgei* wood is an ortholog of NtLIM, which is expressed at a much lower level in *B. burbidgei* than in *P. trichocarpa* (Table S5). In *Nicotiana*, this gene regulates the phenylpropanoid pathway (Kawaoka & Ebinuma, 2001). The first steps of this pathway are shared with the lignin biosynthetic pathway, but expression levels for biosynthetic enzymes from this level of the pathway are not downregulated in the *B. burbidgei* sample in comparison to *P. trichocarpa*.

A final set of regulators worth examining are downstream targets of *MOL1* and *RUL1* (Agusti *et al*., 2011). Though not significantly differentially expressed themselves, three of their targets are upregulated in *B. burbidgei* samples relative to *P. trichocarpa* suggesting this pathway too may be involved in the production of secondarily woody stems.

Comparing wood regulation in *Begonia* with *Populus*

Our *B. burbidgei* transcriptomes show lower expression of biosynthetic genes leading to lignin compared to the situation in *Populus*. This is to be expected, because apart from the fact that *Populus* has primary woodiness and *Begonia* secondary woodiness, the wood anatomy of *Begonia* and *Populus* is very different. For instance, fibre walls in *Populus* wood are more heavily lignified than in *Begonia* wood, and *Begonia* has much more unlignified parenchyma in its wood (especially tall rays) compared with *Populus* (Figs 2, 4, S1). Some of the differences between *Populus* and *Begonia* lignification could be due to sequence variation in lignin biosynthetic genes such as that identified in an ortholog of caffeoyl coenzyme A O-methyltransferase 1 by PAML analysis (Table S4). The differential expression analysis also suggests potential differences in lipid and terpenoid biochemistry in *Begonia* wood (Table S3).

Fasciclin-like arabinogalactans have been implicated in shoot growth (Johnson *et al*., 2011), fibre extension (Huang *et al*., 2013; Liu *et al*., 2013) and secondary wall synthesis and wood formation (reviewed in MacMillan *et al*., 2010). They are thought to act through modification to the cell wall's elasticity. This could also contribute to the differences seen between *Begonia* and *Populus* wood. Orthologs of some FASCICLIN-like arabinogalactans are differentially expressed in wood and cambium stages of *B. burbidgei* stems (Table S3). Other arabinogalactans are found in *B. burbidgei* but not in the genome of herbaceous *B. conchifolia* (Table S5), and some are not recovered from *Begonia* samples though they are expressed at high levels in *Populus* cambium (Table S7). Further analysis of this group of proteins may reveal interesting differences between the formation of wood between *Begonia* and *Populus*.

Potential abiotic variables triggering secondary woodiness in *Begonia*

A comprehensive explanation why secondary woodiness has evolved across *Begonia* is not possible from this initial analysis, but it is clear that a complex mix of different factors is involved. Our Bornean collections point to a strong relationship between increased woodiness, altitude and soil type. All our woody *Begonia* collections were found in montane areas between 1800–2900 m above sea level, and several of these grow on ultramafic rocks, which is also

confirmed for the two known populations of *B. vaccinioides*, the woodiest species of Borneo (Rimi Repin, pers. comm.), and for *B burbidgei* collected along the mountain trail of Mount Kinabalu at 2870 m asl (Lens and Tisun 51, Fig. 4A). Ultramafic rocks and the derived serpentine soils are edaphically stressful for plant growth due to their nutrient deficiencies (especially Ca), low water-holding capacity, and high levels of heavy metals and Mg (Brady, Kruckeberg & Bradshaw, 2005). The combination of poisonous soils and low water-holding capacity must cause stress to the plant, which could lead to wood formation in analogy to the drought stress hypothesis. However, the flora native to serpentine soils (in e.g. California) is mainly composed of herbaceous or primarily woody genera, indicating that only these specific soil conditions are not sufficient to trigger wood formation (Anacker *et al*., 2010; Anacker & Harrison, 2012).

In conclusion, woody *Begonia* species are derived from herbaceous relatives and this trait has evolved independently numerous times across *Begonia*. The selective pressures behind these events remain unclear. Precipitation, altitude and soil conditions may play a role, but it seems more likely that a mix of these factors combined with other abiotic and environmental cues can promote the expression of the wood pathway in *Begonia*. Novel gene expression features of *Begonia* wood in comparison with *Populus* wood (changes in the lignin biosynthetic pathway, changes in FASCICLIN-like arabinogalactans) may be related to the differences in wood anatomy observed. This variation suggests lines of research to better understand the different properties of the wood of primarily and secondarily woody species. The transcriptomes we have generated also provide a number of candidate genes for regulation of wood formation in *Begonia burbidgei* through modification of meristematic activity and cellular anatomy which are worth investigating in other secondarily woody species both in *Begonia* and other genera. The genus *Begonia* offers an excellent opportunity to test the importance of these candidate genes through examination of expression and sequence variation in phylogenetically matched sets of woody and herbaceous species pairs.

ACKNOWLEDGEMENTS

The authors thank Naturalis Biodiversity Center and Sabah Parks for organising and financially supporting the Mount Kinabalu & Crocker Range expedition. We also thank Rimi Repin, Mark Tebbitt and Marc Sosef for their valuable discussions, and Ruth Kiew for identifying the collections from Kinabalu and Crocker Range. Marcelo Pace is acknowledged for collecting the wood sample of *Begonia fruticosa*, and Sukaibin Sumail for sending leaf material of *B. vaccinioides*. Frederic Lens received support from the Alberta Mennega Foundation. We would like to acknowledge the very useful discussions and presentations at the April 2014 Linnean Society meeting 'Collection-based research in the genomic era', which greatly aided our analysis of this data.

REFERENCES

Agusti J, Herold S, Schwarz M, Sanchez P, Ljung K, Dun EA, Brewer PB, Beveridge CA, Sieberer T, Sehr EM, Greb T. 2011. Strigolactone signaling is required for auxin-dependent stimulation of secondary growth in plants. *Proceedings of the National Academy of Sciences USA* **108:** 20242–20247.

Anacker BL, Harrison SP. 2012. Climate and the evolution of serpentine endemism in California. *Evolutionary Ecology* **26:** 1011–1023.

Anacker BL, Whittall JB, Goldberg EE, Harrison SP. 2010. Origins and consequences of serpentine endemism in the California flora. *Evolution* **65:** 365–376.

Andersson-Gunneråas S, Mellerowicz EJ, Love J, Segerman B, Ohmiya Y, Coutinho PM, Nilsson P, Henrissat B, Moritz T, Sundberg B. 2006. Biosynthesis of cellulose-enriched tension wood in *Populus*: global analysis of transcripts and metabolites identifies biochemical and developmental regulators in secondary wall biosynthesis. *The Plant Journal* **45:** 144–165.

Beaman JH, Anderson C, Beaman RS. 2001. *The plants of Mount Kinabalu 4: dicotyledon families Acanthaceae to Lythraceae*. Kota Kinabalu: Natural History Publications.

Blankenberg D, Gordon A, Von Kuster G, Coraor N, Taylor J, Nekrutenko A, Galaxy Team. 2010. Manipulation of FASTQ data with Galaxy. *Bioinformatics* **26:** 1783–1785.

Böhle UR, Hilger HH, Martin WF. 1996. Island colonization and evolution of the insular woody habit in *Echium* L. (Boraginaceae). *Proceedings of the National Academy of Sciences of the United States of America* **93:** 11740–11745.

Bond WJ. 1989. The tortoise and the hare: ecology of angiosperm dominance and gymnosperm persistance. *Biological Journal of the Linnean Society* **36:** 227–249.

Brady KU, Kruckeberg AR, Bradshaw HD Jr. 2005. Evolutionary ecology of plant adaptation to serpentine soils. *Annual Review of Ecology, Evolution, and Systematics* **36:** 243–266.

Brennan AC, Bridgett S, Shaukat Ali M, Harrison N, Matthews A, Pellicer J, Twyford AD, Kidner CA. 2012. Genomic resources for evolutionary studies in the large, diverse, tropical genus, *Begonia*. *Tropical Plant Biology* **5:** 261–276.

Brown JM, Hedtke SM, Lemmon AR, Lemmon EM. 2010. When trees grow too long: investigating the causes of highly inaccurate Bayesian branch length estimates. *Systematic Biology* **59:** 145–161.

Carlquist S. 1974. *Island biology, chapter 10: insular woodiness*. New York: Columbia University Press, 350–428.

Carlquist S. 1985. Wood anatomy of Begoniaceae, with comments on raylessness, paedomorphosis, relationships, vessel diameter, and ecology. *Bulletin of the Torrey Botanical Club* **112:** 59–69.

Carlquist S. 2009. Xylem heterochrony: an unappreciated key to angiosperm origins and diversifications. *Botanical Journal of the Linnean Society* **160:** 26–65.

Carlquist S. 2012. How wood evolves: a new synthesis. *Botany* **90:** 901–940.

Darwin C. 1859. *On the origin of species by means of natural selection*. London: John Murray.

Doorenbos J, Sosef MSM, de Wilde JJFE. 1998. The sections of *Begonia*. *Wageningen Agricultural Papers* **98:** 1–266.

Drummond AJ, Ashton B, Buxton S, Cheung M, Cooper A, Heled J, Kearse M, Moir R, Stones-Havas S, Sturrock S, Thierer T, Wilson A. 2010. Geneious v5.1. Available at: http://www.geneious.com.

Du Z, Zhou X, Ling Y, Zhang Z, Su Z. 2010. AgriGO: a GO analysis toolkit for the agricultural community. *Nucleic Acids Research* **38:** W64–W70.

Dulin MW, Kirchoff BK. 2010. Paedomorphosis, secondary woodiness, and insular woodiness in plants. *Botanical Review* **76:** 405–490.

Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5:** 113.

Forrest LL, Hughes M, Hollingsworth PM. 2005. A Phylogeny of *Begonia* using nuclear ribosomal sequence data and morphological characters. *Systematic Botany* **30:** 671–682.

Givnish TJ. 1998. Adaptive plant evolution on islands: classical patterns, molecular data, new insights. In: Grant PR, ed. *Evolution on Islands*. Oxford: Oxford University Press, 281–304.

Goicoechea M, Lacombe E, Legay S, Mihaljevic S, Rech P, Jauneau A, Lapierre C, Pollet B, Verhaegen D, Chaubet-Gigot N, Grima-Pettenati J. 2005. EgMYB2, a new transcriptional activator from *Eucalyptus* xylem, regulates secondary cell wall formation and lignin biosynthesis. *The Plant Journal* **43:** 553–567.

Goodall-Copestake WP, Pérez-Espona S, Harris DJ, Hollingsworth PM. 2010. The early evolution of the megadiverse genus *Begonia* (Begoniaceae) inferred from organelle DNA phylogenies. *Biological Journal of the Linnean Society* **101:** 243–250.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, Macmanes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, Leduc RD, Friedman N, Regev A. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols* **8:** 1494–1512.

Hamann TD, Smets E, Lens F. 2011. A comparison of paraffin and resin-based techniques used in bark anatomy. *Taxon* **60:** 841–851.

Harrison N. 2012. Phylogenetics of *Begonia* section *Gireoudiana*. D. Phil. Thesis, University of Edinburgh.

Hoch G, Richter A, Körner C. 2003. Non-structural carbon compounds in temperate forest trees. *Plant Cell & Environment* **26:** 1067–1081.

Huang GQ, Gong SY, Xu WL, Li W, Li P, Zhang CJ, Li DD, Zheng Y, Li FG, Li XB. 2013. A fasciclin-like arabinogalactan protein, GhFLA1, is involved in fiber initiation and elongation of cotton. *Plant Physiology* **161:** 1278–1290.

Hughes M. 2008. *An annotated checklist of Southeast Asian Begonia*. Edinburgh: Royal Botanic Garden Edinburgh.

Hughes M, Pullan M. 2007. Southeast Asian *Begonia* database. Available at: http://www.rbge.org.uk

Johnson KL, Kibble NA, Bacic A, Schultz CJ. 2011. A fasciclin-like arabinogalactan-protein (FLA) mutant of *Arabidopsis thaliana, fla1*, shows defects in shoot regeneration. *PLoS ONE* **6:** e25154.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30:** 772–780.

Kawaoka A, Ebinuma H. 2001. Transcriptional control of lignin biosynthesis by tobacco LIM protein. *Phytochemistry* **57:** 1149–1157.

Kiew R. 1998. The unique elements of the limestone flora of Batu Tengar cave (Segarong), Sabah, Malaysia. *Gardens' Bulletin Singapore* **50:** 158–196.

Kiew R. 2001. The limestone begonias of Sabah, Borneo-flagship species for Conservation. *Gardens' Bulletin Singapore* **53:** 241–286.

Langmead B, Salzberg S. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9:** 357–359.

Lee YS. 1974. A study of stem anatomy in *Begonia* L. *Phytologia* **27:** 464–489.

Lens F, Davin N, Smets E, del Arco M. 2013a. Insular woodiness on the Canary Islands: a remarkable case study of parallel evolution. *International Journal of Plant Sciences* **174:** 992–1013.

Lens F, Eeckhout S, Zwartjes R, Smets E, Janssens S. 2012a. The multiple fuzzy origins of woodiness within Balsaminaceae using an integrated approach. Where do we draw the line? *Annals of Botany* **109:** 783–799.

Lens F, Groeninckx I, Smets E, Dessein S. 2009. Woodiness within the Spermacoceae–Knoxieae alliance (Rubiaceae): retention of the basal woody condition in Rubiaceae or recent innovation? *Annals of Botany* **103:** 1049–1064.

Lens F, Schönenberger J, Baas P, Jansen S, Smets E. 2007. The role of wood anatomy in phylogeny reconstruction of Ericales. *Cladistics* **23:** 229–254.

Lens F, Smets E, Melzer S. 2012b. Stem anatomy supports *Arabidopsis thaliana* as a model for insular woodiness. *New Phytologist* **193:** 12–17.

Lens F, Tixier A, Cochard H, Sperry JS, Jansen S, Herbette S. 2013b. Embolism resistance as a key mechanism to understand plant adaptive strategies. *Current Opinion in Plant Biology* **16:** 287–292.

Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12:** 323.

Liu H, Shi R, Wang X, Pan Y, Li Z, Yang X, Zhang G, Ma Z. 2013. Characterization and expression analysis of a fiber differentially expressed Fasciclin-like arabinogalactan protein gene in Sea Island cotton fibers. *PLoS ONE* **8:** e70185.

Liu L, Missirian V, Zinkgraf M, Groover A, Filkov V. 2014. Evaluation of experimental design and computational parameter choices affecting analyses of ChIP-seq and RNA-seq data in undomesticated poplar trees. *BMC Genomics* **15** (Suppl 5):S3.

MacMillan CP, Mansfield SD, Stachurski ZH, Evans R, Southerton SG. 2010. Fasciclin-like arabinogalactan proteins: specialization for stem biomechanics and cell wall architecture in *Arabidopsis* and *Eucalyptus*. *The Plant Journal* **62:** 689–703.

Marshall DC. 2010. Cryptic failure of partitioned Bayesian phylogenetic analyses: lost in the land of long trees. *Systematic Biology* **59:** 108–117.

Melzer S, Lens F, Gennen J, Vanneste S, Rhode A, Beeckman T. 2008. Flowering time genes modulate meristem determinacy and growth form in Arabidopsis. *Nature Genetics* **40:** 1489–1492.

Miller MA, Pfeiffer W, Schwartz T. 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In: Proceedings of the Gateway Computing Environments Workshop (GCE). New Orleans: IEEE, 1–8.

Moonlight PW, Richardson JE, Tebbitt MC, Thomas DC, Hollands R, Hughes M. 2015. Continental-scale diversification patterns in a mega-diverse genus: the biogeography of Neotropical *Begonia* L. *Journal of Biogeography* (in press).

Moyers BT, Rieseberg L. 2013. Divergence in gene expression in uncoupled from divergence in coding sequence in a secondarily woody sunflower. *International Journal of Plant Sciences* **174:** 1079–1089.

Nylander JAA. 2004. MrModeltest v2.3. Available at: http://www.abc.se/~nylander.

Nylander JAA, Wilgenbusch JC, Warren DL, Swofford DL. 2008. AWTY (Are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. *Bioinformatics* **24:** 581–583.

Plana V. 2003. Phylogenetic relationships of the Afro-Malagasy members of the large genus *Begonia* inferred from *trn*L intron sequences. *Systematic Botany* **28:** 693–704.

Plana V, Gascoigne A, Forrest LL, Harris D, Pennington RT. 2004. Pleistocene and pre-Pleistocene Begonia speciation in Africa. *Molecular Phylogenetics and Evolution* **31:** 449–461.

Plana V, Sands MJS, Beentje HJ. 2006. Begoniaceae. In: Beentje HJ, Ghanzafar SA, eds. *Flora of tropical East Africa*. Kew: Royal Botanic Gardens, 1–53.

Prunet N, Morel P, Thierry A-M, Eshed Y, Bowman JL, Negrutiu I, Trehin C. 2008. REBELOTE, SQUINT, and ULTRAPETALA1 function redundantly in the temporal regulation of floral meristem termination in *Arabidopsis thaliana*. *The Plant Cell* **20:** 901–919.

Rambaut A, Drummond AJ. 2009. Tracer v1.5. Available at: http://beast.bio.ed.ac.uk/Tracer.

Reitsma JM. 1984. *Begonia* section *Bacca begonia* Reitsma, sect. nov. *Agricultural University Wageningen Papers* **84:** 95–111.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European molecular biology open software suite. *Trends in Genetics* **16:** 276–277.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26:** 139–140.

Robischon M, Du J, Miura E, Groover A. 2012. The *Populus* class III HD ZIP, popREVOLUTA, influences cambium initiation and patterning of woody stems. *Plant Physiology* **155:** 1214–1225.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19:** 1572–1574.

Sang J, Kiew R, Geri C. 2013. Revision of *Begonia* (Begoniaceae) from the Melinau Limestone in Gunung Mulu National Park and Gunung Buda National Park, Sarawak, Borneo, including thirteen new species. *Phytotaxa* **99:** 1–34.

Thomas DC, Hughes M, Phutthai T, Ardi WH, Rajbhandary S, Rubite R, Twyford AD, Richardson JE. 2012. West to east dispersal and subsequent rapid diversification of the mega-diverse genus *Begonia* (Begoniaceae) in the Malesian archipelago. *Journal of Biogeography* **39:** 98–113.

Thomas DC, Hughes M, Phutthai T, Rajbhandary S, Rubite R, Ardi WH, Richardson JE. 2011. A non-coding plastid DNA phylogeny of Asian *Begonia* (Begoniaceae): evidence for morphological homoplasy and sectional polyphyly. *Molecular Phylogenetics and Evolution* **60:** 428–444.

Tuskan GA, DiFazin S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, Schein J, Sterck L, Aerts A, Bhalerao RR, Bhalerao RP, Blaudez D, Boerjan W, Brun A, Brunner A, Busov V, Campbell M, Carlson J, Chalot M, Chapman J, Chen GL, Cooper D, Coutinho PM, Couturier J, Covert S, Cronk Q, Cunningham R, Davis J, Degroeve S, Déjardin A, Depamphilis C, Detter J, Dirks B, Dubchak I, Duplessis S, Ehlting J, Ellis B, Gendler K, Goodstein D, Gribskov M, Grimwood J, Groover A, Gunter L, Hamberger B, Heinze B, Helariutta Y, Henrissat B, Holligan D, Holt R, Huang W, Islam-Faridi N, Jones S, Jones-Rhoades M, Jorgensen R, Joshi C, Kangasjärvi J, Karlsson J, Kelleher C, Kirkpatrick R, Kirst M, Kohler A, Kalluri U, Larimer F, Leebens-Mack J, Leplé JC, Locascio P, Lou Y, Lucas S, Martin F, Montanini B, Napoli C, Nelson DR, Nelson C, Nieminen K, Nilsson O, Pereda V, Peter G, Philippe R, Pilate G, Poliakov A, Razumovskaya J, Richardson P, Rinaldi C, Ritland K, Rouzé P, Ryaboy D, Schmutz J, Schrader J, Segerman B, Shin H, Siddiqui A, Sterky F, Terry A, Tsai CJ, Uberbacher E, Unneberg P, Vahala J, Wall K, Wessler S, Yang G, Yin T, Douglas C, Marra M, Sandberg G, Van de Peer Y, Rokhsar D. 2006. The genome of black

cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313:** 1596–1604.

Vanholme R, Cesarino I, Rataj K, Xiao Y, Sundin L, Goeminne G, Kim H, Cross J, Morreel K, Araujo P, Welsh L, Haustraete J, McClellan C, Vanholme B, Ralph J, Simpson GG, Halpin C, Boerjan W. 2013. Caffeoyl shikimate esterase (CSE) is an enzyme in the lignin biosynthetic pathway in *Arabidopsis*. *Science* **341:** 1103–1106.

Wenkel S, Emery J, Hou BH, Evans MM, Barton MK. 2007. A feedback regulatory module formed by LITTLE ZIPPER and HD-ZIPIII genes. *The Plant Cell* **19:** 3379–3390.

Wernersson R. 2006. Virtual ribosome–a comprehensive DNA translation tool with support for integration of sequence feature annotation. *Nucleic Acids Research* **34:** W385–W388.

Wernersson R, Pedersen AG. 2003. RevTrans: multiple alignment of coding DNA from aligned amino acid sequences. *Nucleic Acids Research* **31:** 3537–3539.

Whittaker RJ, Fernández-Palacios JM. 2007. *Island biogeography*. *Ecology, evolution, and conservation, 2nd edn*. New York: Oxford University Press.

Wu Z-J, Li X-H, Liu Z-W, Xu Z-S, Zhuang J. 2014. De novo assembly and transcriptome characterization: novel insights into catechins biosynthesis in *Camellia sinensis*. *Plant Biology* **14:** 277.

Xu W, Li R, Zhang N, Ma F, Jiao Y, Wang Z. 2014. Transcriptome profiling of *Vitis amurensis*, an extremely cold-tolerant Chinese wild *Vitis* species, reveals candidate genes and events that potentially connected to cold stress. *Plant Molecular Biology* **86:** 527–541.

Yang Z. 2007. PAML4 Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24:** 1586–1591.

Zhang J, Nieminen K, Alonso Serra JA, Helariutta Y. 2014. The formation of wood and its control. *Current Opinion in Plant Biology* **17:** 56–63.

Zhu S, Tang S, Tang Q, Liu T. 2014. Genome-wide transcriptional changes of ramie (*Boehmeria nivea* L. Gaud) in response to root-lesion nematode infection. *Gene* **552:** 67–74.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Figure S1. Transverse (A) and tangential (B) wood section of *Populus trichocarpa* showing more lignified tissue, smaller alternate intervessel pitting (long arrows), and narrower and shorter rays (short arrows) compared to *Begonia burbidgei*.

Table S1. List of specimens used in this paper, with reference to their voucher data, place of origin and habitat. Voucher material is deposited in Sabah Parks (Sabah, Malaysia) and Forest Research Institute Malaysia. Species with an asterisk are woody.

Table S2. GenBank accession numbers. GenBank accession numbers in bold font indicate sequences newly generated for this study. All other sequences were downloaded from the nucleotide database of the National Centre for Biotechnology Information (http://www.ncbi.nlm.nih.gov/).

Table S3. Sequences differentially expressed between cambium stage and wood stage *Begonia burbidgei* samples.

Table S4. Sequences with high divergence between woody *Begonia burbidgei* and herbaceous *Begonia venustra*.

Table S5. *Begonia burbidgei* sequences with matches in *Populus trichocarpa* but no matches in the genome of herbaceous *Begonia conchifolia*.

Table S6. Sequences with differential expression between *Begonia burbidgei* stems and *Populus trichocarpa* cambium.

Table S7. Sequences present in a transcriptome from *Populus trichocarpa* cambium but not in the *Begonia burbidgei* transcriptome.

Table S8. Relative expression levels for each enzyme class between *Begonia burbidgei* and *Populus trichocarpa*.